汽车千车故障数的预测与分析

郑晓练，管河山，陈捷

指导教师：谭忠

（厦门大学 福建 厦门 361005）

摘要：针对原有千车故障数统计方法上的不足，本文从改进统计方法着手，提出一种新的统计方法即重新定义千车故障数，然后利用数据挖掘中的聚类分析方法将具有相同特征的批次综合起来考虑，建立通用的运筹模型。针对缺失数据、近期预测这两个问题，本文对通用模型进行调整，“学习”出同类数据间的不同权值，然后利用加权数据，并通过拟合曲线来求出预测值。由于远期预测中数据的严重缺乏，则是从纯粹统计学的角度出发,计算得到预测值。预测模型通用性强，适用面较广。本文应用了 SAS 和 MATLAB 两种软件来求解上述模型，预测结果准确率较高，并且符合实际情况。

关键字：聚类分析、曲线拟合、权值学习、SAS、MATLAB

1．问题的提出（略）

2．对统计方法的改进

整车或某个部件的“千车故障数”常用于描述轿车的质量。它原先的定义为：将轿车按生产批次划分成若干个不同的集合，每个集合中迄今已售出的全部轿车中,在相同使用时间长度内的整车或某个部件的保修总次数乘以 1000 再除以迄今已售出的轿车数量。这样的统计方法下，数据具有明显的时滞性，部分数据的误差很大。由于所用千车故障数的统计是达到相同使用时间长度（如 i 月）的故障数与总共的销售量的比，分母中包含了已经到达和未达到该时间长度的总共车辆，显然与要求的千车故障率相比相差了未到达该使用时间常数的零件数的统计值，即样本空间不同。

为了避免以上不足给预测造成的影响，我们可以由当前的数据推出另一套数据:

$A_j=(\sum_{m=0}^{j}\sum_{i=0}^{N}x_ip_m/\sum_{i=0}^{N}x_i)\times1000$；$A_j^{'}=(\sum_{m=0}^{j}\sum_{i=0}^{N-j}x_ip_m/\sum_{i=0}^{N-j}x_i)\times1000$；$\overline{A_j}=(\sum_{m=0}^{j}\sum_{i=0}^{\infty}x_ip_m/\sum_{i=0}^{\infty}x_i)\times1000$；

其中 A_j 为表中给出的某批次零件使用了 j 月的千车故障数，A_j’是只计算达到该使用年限的车的千车故障数，$\overline{A_j}$ 是千车故障数的期望值。x_i 为车的月销售量，p_m 为该批次第 m 个月的故障率， N 为车卖出的月份数。

使用年限由于 A_j 和 A_j’两者的期望均趋近于 $\overline{A_j}$，因此它们都是无偏估计。但是可以看出 A_j’为一致无偏估计，而 A_j 不是一致的，因此 A_j’比 A_j 更快趋近期望 $\overline{A_j}$，A_j’对于预测将是更理想的数据。从另一方面解释，相对与 A_j，A_j’剔除了销售量中使用时间没达到 j 月的车辆，更符合千车故障率统计的要求。

同时根据零件寿命的分布，可以得出 p_m 服从指数分布，其 λ 值在 A_j’和 A_j 的表达式中相同。

我们有充足的理由提出一种新的统计方法即重新定义所谓的千车故障数，将之定义为：相同使用长度的车保修总次数与在时间上能够产生这些保修次数的车的比值乘以 1000，这样就可以使当前的千车故障数更具可靠性，能更快地趋近于千车故障数的期望值，这样的数据对于预测将更有意义。

其表达式为 $A_j^{'}=(\sum_{m=0}^{j}\sum_{i=0}^{N-j}x_ip_m/\sum_{i=0}^{N-i}x_i)\times1000$。

3．预测模型假设与符号说明

Ⅰ模型假设

1） 车在每个月短时间内售出，不考虑月内的销售量分布。
2） 车在每个月的销售量服从在统计时间内的均匀分布。
3） 同一批次产的零件或整车质量相同。
4） 顾客车一坏就送到修配厂修理,不存在滞留的情况。
5） 各维修部门返馈数据及时,即维修情况能准时返送到总部数据中心。
6） 质量指标λ包含了所有其他的随机影响，即所有随机影响都体现在λ的变化中

Ⅱ符号说明

$A_{k,j}$：表中给出的第 k 批次零件使用了 j 月的千车故障数

$A_{k,j}'$：还原的千车故障数

S_{total}：某一批次的总销售量

$0x_i$：车的月销售量

p_m：某批次第 m 个月的故障率

time：制表时间

N：车卖出到制表时间的月数

j：车使用月数

quan(k)=第 k 个批次在加权时的劝值

4．问题分析

Ⅰ数据还原

我们将表中的数据按新的方法进行还原，使之合理，还原方法为以上提出的新的千车故障数的求法，即：$A_{k,j}' = (\sum_{m=0}^{j}\sum_{i=0}^{N-j} x_i p_m / \sum_{i=0}^{N-i} x_i)\times 1000$。

但是考虑到 p_m 的分布是服从指数分布,为 $p_m = 1-e^{-\lambda m}$，λ值的求得本来就必须通过给出的 A_j 来拟合，求出最优值，本身就存在一定的误差。同时，可以看到每个时间长度下的故障数对于销售量来说是很少的，这部分数值的小变动不会对还原数据产生太大的影响，我们假设销售量服从均匀分配，因此我们将以上的表达式-简化为：

$$A_{k,j}' = \frac{A_{k,j} * S_{total}}{S_{total} - S_{total} * (j-1)/(time-k)} / 1000 = \frac{A_{k,j} * (time-k)}{(time-k-j+1)} / 1000$$

其中 time-k 表示第 k 批次到制表时间所间隔的月份数。分母项的是全部的销售量减去后 j-1 个月的销售量，含义即为使用时间长度>j 的销售车数。

Ⅱ数据分析

首先考虑经过处理后数据的意义，$A_{k,j}'$是表示某批车中使用了 j 个月的故障零件数占总车数的比例，也就等于零件的寿命≤j 月的概率。既 $A_{k,j}' = F(x \le t) = 1-e^{-\lambda * t}$，$A_{k,j}'$服从指数分配。

其次，我们画出还原数据的表中每个批次的千车故障数的曲线图，每一批次画一条曲线，线上横坐标为 j 点表示该批次在第 j 月的千车故障数，见图 1。可以看到：第一，总体曲线呈指数分布，说明以上的数据分析是准确的。第二：某些批次的曲线比较类似，λ值相似。考虑到如果单独对某批次的千车故障数进行处理，其中包含的随机误差就比较大，但是如果将具有相同特征的批次综合起来考虑，就可以有效地消除大量的随机误差。因此，我们利用了数据挖掘技术对表中数据进行了聚类分析。

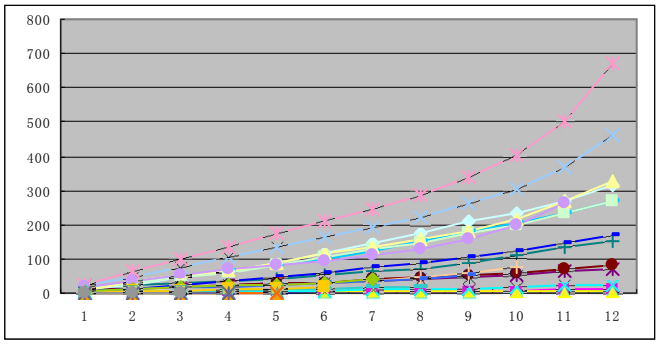

图 1

Ⅲ聚类分析

聚类分析是一种新兴的多元统计方法，是当代分类学与多元分析的结合。聚类分析是将分类对象置于一个多维空间中，按照它们空间关系的亲疏程度进行分类，也就是根据事物彼此不同的特征进行辨认，将具有“相似”特征的事物聚为一类，使得同一类的事物具有高度的相似性。聚类的方法很多，有：最短距离法、最长距离法、中间距离法等。我们采用的聚类方法是最短距离法，主要步骤分四步：

1. 根据研究的具体目的来选择合适的聚类变量，我们要预测的是千车故障数，因此选择 $A_{k,j}'$ 来作为聚类变量。
2. 计算相似性测度，根据现在我们选择的样本聚类，选择相应的距离计算公式计算类与类之间的距离。
3. 再次选定聚类方法进行聚类，我们选择的是用最短距离法。
4. 对结果进行验证，根据最后分的类数以及一个类里各个样本之间的相似关系来判断最后的结果是否合理，不合理就再选择类数等参数进行分类。分类的类数以及数据是否合理是根据研究目的进行判断的，不同研究目的分类不同。

我们使用 SAS 软件进行聚类分析，方法选择用 method 参数确定，最后类数由参数 rsq，pseudo，ccc 等判定，再结合实际目的决定最终分类数。

5．模型建立

0205 批使用月数为 18 的数据预测代表了对缺失数据的预测，0306 批使用月数为 9 的数据预测代表了对近期数据的预测，而 0310 批使用月数为 12 的数据预测代表了对远期数据的预测。由于对于预测点 1，2，数据可以认为是足够的，因此都可以用现有表中数据预测出，而预测点 3 即 0310 批 j=12 的 $A_{k,j}'$ 的预测中，数据是严重缺乏及不合理的，对其的预测方法我们将最后讨论。对于预测点 1，2 来说，我们是在基本模型的基础上针对不同种数据的预测做了不同的调整，使该类数据的预测可以更准确。

由于我们假设同一批次的零件质量相同，即决定故障数分布质量指标是相同的，反映到表达式上，为 λ 相同。因此所有批次通用的模型为：$A_{k,j}'=1-e^{-\lambda * t} \mid t=j$

首先要对同一类的数值进行加权平均，权值由以下模型取最忧，这样可以列出以下运筹模型。其中 $A_{m,j}$ 为加权数据中使用时间 j 个月的千车故障数，$A_{f,j}$ 为预测批次中使用时间 j 个月的千车故障数，n 为几个批次共有的与使用时间长度相对应的数据集数。

$$\text{MIN} \sum_{j=1}^{n} \left(A_{m,j}-A_{f,j}\right)^2$$

$$\text{S.T.}\quad \begin{cases} A_{m,j}=\sum_{i=1}^{l} A'_{ij}\times\alpha_i \\ A_{m,j}=1-e^{-\lambda' t}\Big|_{t=j} \\ \sum_{i=1}^{n}\alpha_i=1 \\ A_{m,j}\geq 0 \end{cases}$$

由这个模型可以算出最优的α_i。

问题一：缺失数据预测模型的建立（0205 批使用月数为 18 的数据预测）

对于该数据预测，可以直接用以上求出的最优权值来求出同类中所有批次的加权平均数据，利用该组数据来拟合λ'，带入 $A_{k,j}'=1-e^{-\lambda * t}\mid_{t=j}$ 这样就可以求得该预测点的数据，其中最优权值的确定是一个不断“学习”的过程。

问题二：短期预测模型的建立（0306 批使用月数为 9 的数据预测）

对于该数据预测，由于在 t>9 情况下给的数据较少，统计该部分数据产生的误差就比较大，因此这个批次的λ值不具有质量指标的特征含义，不可以直接拟合λ。对于这类数据的预测，我们将与这个批次一类的几个批次的 $A_{k,j}'$ 先做加权平均，目标是与 0306 这个批次在相同使用时间长度上的 $A_{k,j}'$ 方差最小（$A_{k,j}'$ 中 k 是定值，j 在变化）。以上建立运筹模型同样可应用于这种情况，与预测第 0205 批数据的不同之处就在于 $A_{f,j}$ 定义不同，权值确定方法不同。0205 批数据预测时，在准确率标准下不断通过前面小量数据来预测后面已经存在数据，从而“学习”出权值，也就是 $A_{f,j}$ 是在不断“学习”被预测的那批已给出的数据，是随着学习过程变化的。而本批数据的预测，$A_{f,j}$ 是定值，就是 0306 批使用月数小于 9 的千车故障数，因此，本批数据的权值可以通过直接求解以上的运筹模型来得到。

求得权值后就可以用加权数据的曲线来逼近预测的曲线。既然它们可以归为一类，即特征相似，说明它们近似服从一个λ分布，因此预测曲线的 λ' 值可以近似用拟合曲线的λ来代替，从而就可求出预测值。

问题三：远期预测模型的建立（0310 批使用月数为 12 的数据预测）

根据题意，该千车故障数的数据表是 2004 年 4 月 1 日整理出来的，而日常售后服务数据的反馈是瞬时的，即厂家可以迅速地更换出最新的统计数据，由此可知，已经给出的轿车部件千车故障数的数据表每列都有两行（最上两行空白）的数据已经统计出来，而在表中并未给出。为此根据各列的独立分布性质，且每行（每批产品）服从指数形式的分布，可以先将原始表中的已经统计出却没有给出的数据预测出来，得到一个新表。将新表中的行（0201-0309）和后 7 列（包括 0 月）这些数据，从统计的角度出发，做出每列的分布图，可以得知他们是近似正态分布，每一列数据是独立的，而且可以说近似的服从同一分布（近似正态分布），从而我们可以利用第 3 列（即使用 12 个月的）近似服从正态分布的性质来预测 0310 批的使用 12 时的千车故障数。从图象可以看出，我们可以用图表中前 4 个点的平均值来预测表中未知的后 4 个月的平均数，而用后 4 个月的平均数来预测 0310 批使用月数 12 时的千车故障数。

由于以上数据的预测是建立在还原后的数据基础上，最后要将数据换算回原始表中定义方式下的数据做为最后的预测值。

6．模型求解

Ⅰ问题一的解（0205 批次使用月数为 18 的千车故障数）

我们运用了 SAS 软件与 MATLAB 软件对模型进行了求解，用 SAS 软件求的聚类分析的结果为 0205 和 0206 两个批次是一类的。

对于“学习”过程的求解为，先试着用尽量小的数据量来拟合出分布曲线，后用这个分布算式预测在此使用月数之后的已知数据，并计算方差；然后逐渐加大用于拟合的数据量，重复以上过程，如先用使用月数 j=1 到 9 的数据来估计 j 为 10 到 12 的 3 个点，然后用 j=1 到 10 的数据预测后 2 个数，最后再用 j=1 到 11 的数据预测第 12 个点，这个过程通过 matlab 程序的模拟，可得到这个类中的最优权值，为 ω(0205)=0.6，ω(0206)=0.4，拟合出 λ 带入曲线在换算成原来表中的数据，则 j=18 这点的千车故障数约为 51.99。

加权平均值图和 0205 批千车故障数图，见图 2：

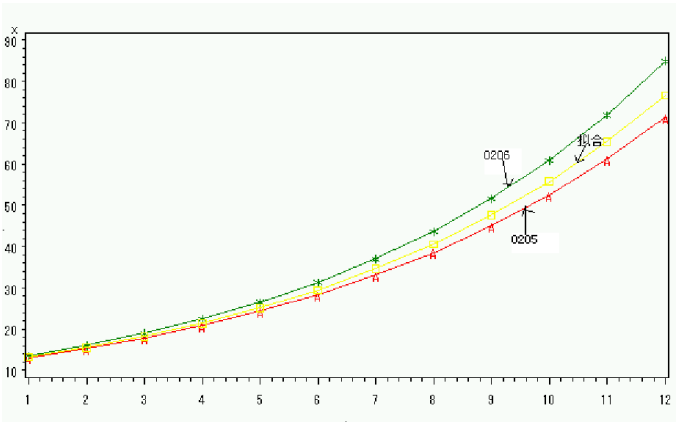

图 2

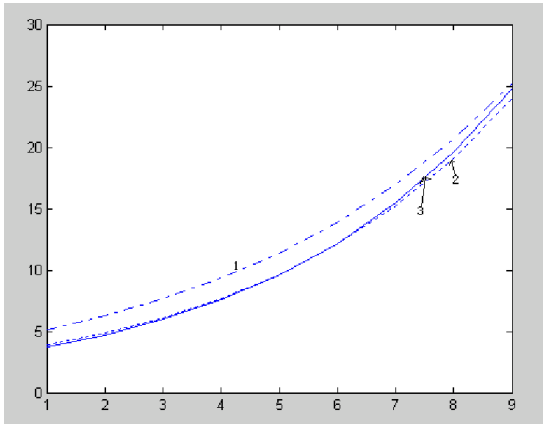

图 3

Ⅱ 问题二的解（0306 批次使用月数为 9 的千车故障数）

本批次零件和 0201，0202，0203，0204，0205，0206，0304，0305 这 8 个批次是可以分做一类的，这类中同样通过求解线形规划模型得到这个类中的最优权值为表 1 所示。

权号 i	1	2	3	4	5	6	7	8
权值 ω（i）	0.227047	0.162224	0.127163	0.105077	0.094944	0.090855	0.091566	0.101124

表 1

其中 i 表示用于求加权平均值的第 i 份数据，这里的排列是按批次时间，由 0201—0305 排列，分别为第 1 份—第 8 份数据。从而先算出加权平均这组数据在使用时间长度为 9 月时的千车故障率 $A'_{m,9}$=24.50471 ，然后拟合出这条曲线的 λ 值，而 0306 批的 λ' 就可以近似用 λ 来代替。换算成原来的表中数据，求出 j=9 时的千车故障数约为 6.87。

产生逼近直线的加权平均数据：

使用月数	9	8	7	6	5	4	3	2	1
千车故障率	24.50471	20.55735	17.57892	14.61866	12.06629	9.955031	7.670813	5.553821	3.242005

表 2

逼近图（图 3）如下：

1— 表示由加权平均数据拟合的曲线

2— 表示由表 1 中给出的 0306 批使用时间长度在 1 到 8 月的数据拟合的曲线

3— 表示由线 2 中的 8 个原始数据点加上预测的使用时间长度为 9 月的数据点共同和的曲线。

由图 3 可以很清楚地看出 3 条线之间的关系，可见线 1，2，3 趋势都基本相同，同一类的关系得到验证，而通过线 1 逼近线 2 产生了线 3，线 3 含有预测量的值，它的千车故障率随使用时间长度的增大，增大的速率超过了全由原始数据点拟合的线 2，这也说明了线 3 被线 1 矫正后，部分克服

了线 2 由于统计数据量太少引起千车故障率较小的不合理情况，其值更符合实际。这样也就证明了我们的预测是比较成功的

Ⅲ问题三的解（0310 批次使用月数为 12 的千车故障数）

根据还原数据表和转换公式，以及以上对该模型的分析

计算结果为：（8.42*12/23+10.38*11/22+7.91*10/21+25.42*9/20）/4=6.17

7．模型评价

以上建立的基本模型主要侧重于还原合理数据和数据的聚类分析,强调在同一类中对数据进行建模处理,以及用于预测的值的可靠性，这样做对预测具有广泛的意义,可适用于各种类型的值的预测,只是根据类型不同求具体值时方法不同而已。而且这样的模型对于实际数据的吻合度高，同时有零件寿命服从指数分布来做为该模型的理论支持，因此这样的模型是可靠的。

对于缺失数据的预测，体现为 0205 批次使用月数为 18 的千车故障数预测，拟合曲线预测的准确率随拟合数据变多而升高，所以用 0205 的全部 12 个数据来拟合曲线，预测的准确率很高。这个批次的预测情况比较好是因为产生该批次千车故障数的总共车数多，距离制表时间长，数据统计相对比较充分，比较可靠。

对于近期值的预测，体现为 0306 批次使用月数为 9 的千车故障数预测，由于可以用于统计的车辆数比较少，因此实际上这些点的千车故障数比表中体现的数据要多，我们用以上的模型对其进行预测后，0306 批千车故障数呈现出的趋势保持了和同类的几个批次的趋势一致，而且千车故障数的数值提高了，因此，这样的预测值更加符合实际值。由此我们可以认为这样的预测方法是比较成功的。

对于远期值的预测，使用月数内的千车故障数互相间是相互独立的，从统计学角度上说，这样的预测是完全合理的。

8．参考文献

[1]《运筹学》教材编写组，运筹学，出版地：清华大学出版社，2002 年。

[2]阮桂海等，SAS 统计分析实用大全，出版地：清华大学出版社，2003 年。

[3]岳朝龙，黄永兴，严忠，SAS 系统与经济统计分析，出版地：中国科技大学出版社，2003 年。

[4]姜启源，数学模型，出版地：高等教育出版社，2001 年。

Prediction and Analysis for the Broken-down Number Per Thousand Cars

ZHENG Xiao-lian,　GUAN He-shan,　CHEN Jie

(Xiamen University, Fujian, Xiamen, 361005)

Abstract　Responding to the drawbacks of the old statistical method of the number of down-broken cars per thousand,this paper attempts to improve the existing statistical method to recreate a new one, namely we redefine the number of down-broken cars per thousand and utilize the cluster analysis in data mining to establish a universal operational model, so as to comprehensively consider those groups which members share certain properties in common.For the questions of scanty of data and short-term forecast,we adjust the universal operational model and study the different weight among the same class.Base on the weighted data,we can get the value of forecast through fitting the curve. Since acute shortage of data in long-term forecast, we start from the view point of pure statistics and get the forecast value. The forecast model can be used comprehensively. This paper solves the model mentioned above by using SAS and MATLAB , the forecasted result with high veracity of and conforms to practical condition.

Keywords　cluster analysis, curve fitting,　weight learning,　SAS,　MATLAB